Title	家系図のネットワーク構造の解析 (第5回生物数学の理論とその応用)
Author(s)	堀内, 陽介; 水口, 毅; 守田, 智
Citation	数理解析研究所講究録 (2009), 1663: 11-13
Issue Date	2009-09
URL	http://hdl.handle.net/2433/141024
Right	
Type	Departmental Bulletin Paper
Textversion	publisher

数理解析研究所講究録
第1663巻 2009年 11-13

家系図のネットワーク構造の解析

Analysis of network structure of family trees

堀内 陽介 (Yosuke Horiuchi), 水口 毅 (Tsuyoshi Mizuguchi) [1]
大阪府立大学大学院工学研究科数理工学分野
Department of Mathematical Sciences, Osaka Prefecture University
守田 智 (Satoru Morita)
静岡大学工学部システム工学
Department of Systems Engineering, Shizuoka University

有性生殖を行う生物個体の先祖数は世代をさかのぼる毎に倍になる．この先祖数の指数関数的増大は，過去のある時代までさかのぼれば先祖数が当時の総人口を超えてしまうというパラドクスを引き起こす．このパラドクスを解消するのは，「先祖個体の中には重複した役割を持つものがいる」という事実であるが，この事実は家系図が絡み合ったネットワーク構造をなしていることも示している．この家系図のネットワーク構造に関して，いくつかの研究がなされてきた [1, 2, 3, 4, 5]．しかし，先行研究は単純化されたモデルに基づいているものが多く，実際の生物のデータによる実証はほとんどない．それは，実際の生物の家系図で長期に渡って保存されているものが少ないからである．本研究では，実際の生物として競走馬に着目し，その家系図データを用い，祖先同士をつなぐネットワークがどのような構造をなしているかを明らかにすることを目指している．競走馬の血統情報は良く保存されており，ユニークに割り当てられた馬名に対して，性別，生年，父馬名，母馬名が検索できる．得られた親の名前を用いて再帰的に検索することで，ある個体の先祖を漏らさずに検索することができる．親馬が「不明」になってしまえばそれ以上さかのぼることができないが，最近の馬からなら平均して17世代（約200年）くらいまでさかのぼることができる．

まず，注目する個体を主個体と呼ぼう．主個体 α および α の直接の先祖全体からなる集団に対し，個体を点で表し，親子関係にある点を線で結んだものを α の木と呼ぶ[2]．上に述べた祖先数のパラドクスは，主個体の木の中に複数回登場する先祖 γ の存在によって解消される．このとき，祖先 γ は二本の木の共通部分と見ることもできる．さらに γ がある二本の木の共通部分であれば γ の先祖も必ず共通部分であることから，任意の二本の木は世代を遡るにつれ共通部分が多くなることが予想される．この過程を二本の木の合体過程と呼ぼう．

Derrida らは，離散化された世代 G 毎に総数 $N(G)$ の個体が全て入れ替わるという単純なモデルに対して，二本の木の合体過程を以下のように特徴づけた [3]．ある主個体の木に属する各先祖に対して，どの程度主個体に影響があるかを特徴づける量として「ウェイト」を以下のように帰納的に定義する：主個体 α に対する先祖 γ のウェイト w_γ^α は，

$$w_\gamma^\alpha = \frac{1}{2} \sum_{\gamma' \in C(\gamma)} w_{\gamma'}^\alpha, \qquad (1)$$

である．ここで，$C(\gamma)$ は，γ の子の集合を表す．主個体自身のウェイトは $w_\alpha^\alpha = 1$ とする．いくつ

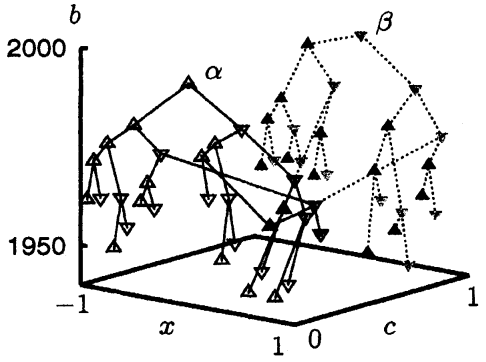

図1．競走馬のデータを用いた二本の木の合体過程．b 軸は生年，x 軸は祖先座標，c 軸は，二主個体のウェイトの比を表す．△（▽）は α の木の雌（雄），▲（▼）は β の木の雌（雄）を表す．

[1]corresponding author: gutchi@ms.osakafu-u.ac.jp
[2]グラフ理論における「木」の定義とは異なる．

か実例をあげると，親のウェイトは 1/2, 祖父母のウェイトは 1/4, 曽祖父母のそれは 1/8 である．すなわち，ウェイトはいわゆる「血の濃さ」であり，世代をさかのぼる毎に 1/2 倍される．ただし，先祖個体が複数の役割を持つ場合は，それぞれの役割のウェイトの和を持つとする．二本の木の合体過程を定量的に特徴づける量として，同時代生まれの二主個体 α, β それぞれの木に属する先祖 γ のウェイト $w_\gamma^\alpha, w_\gamma^\beta$ の関係に着目しよう．γ をちょうど G 世代前のすべての個体（個体数 $N(G)$ ）とし，ウェイト $\{w_\gamma^\alpha\}, \{w_\gamma^\beta\}$ を $N(G)$ 個の成分を持つベクトルとみなせば，内積 $Y^{\alpha,\beta} \equiv \sum_\gamma w_\gamma^\alpha \cdot w_\gamma^\beta$ をノルム $X^\alpha \equiv \sqrt{\sum_\gamma (w_\gamma^\alpha)^2}$ で正規化した量

$$q^{\alpha,\beta}(G) \equiv \frac{Y^{\alpha,\beta}}{X^\alpha \cdot X^\beta} \tag{2}$$

は，二つのベクトルのなす角の余弦であり，両者が完全に一致すれば 1, 共通部分が全く無ければ 0 になる．このことから分かる通り，$q^{\alpha,\beta}(G)$ は G 世代前の木の重複の程度を示す量になっている．たとえば，α, β として（両親とも同じ）兄弟姉妹をとれば，$G > 0$ に対して $q^{\alpha,\beta}(G) = 1$ となり，両者の木は 1 世代さかのぼることで完全に一致する．この $q^{\alpha,\beta}(G)$ の集団平均 $\langle q(G) \rangle$ の G 依存性が理論的に見積もられている．

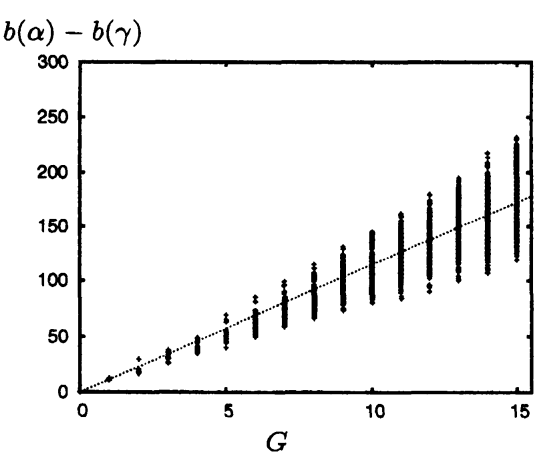

図 2. 役割毎の主個体と先祖の生年差と世代．縦軸は主個体と先祖の生年差，横軸は世代．たとえば，祖父 ($G = 2$) が，主個体より 20 年前に生まれている場合，$(2, 20)$ にプロットされる．同一先祖が複数の役割をこなしている場合は，全ての役割についてプロットする．シンボルは 1996 年生まれの中からランダムに選んだ 1 頭を主個体としたもの．点線は，ランダムに選んだ 46 頭分のデータを原点を通る直線でフィットしたもの．その傾きは $a \approx 11.47$ である．

この理論は競走馬にたいしてどの程度妥当なのだろうか．図 1 は，競走馬のデータを用いた二本の木の合体過程の典型例である．1996 年生まれの二頭 α, β を主個体とし，4 世代までの木を b 軸に生年，x 軸に祖先座標[3] をとり，c 軸に二主個体のウェイトの比 $w_\gamma^\beta/(w_\gamma^\alpha + w_\gamma^\beta)$ を描いたものである．これをみると，たとえば α の父の父と β の母の父の父が一致していることが分かる．4 世代まででは共通祖先の数は少ないが，世代をさかのぼるにしたがって共通祖先の数も増大する．

競走馬から得られたデータと前述の理論の結果とを比較するためには，すこし工夫しなければならない．というのも，理論に使用された単純なモデルにはいくつかの仮定があり，その中には実際の生物の家系図には当てはまらないものもあるからである．たとえば，世代毎に全ての個体が入れ替わるという仮定は，競走馬には当てはまらない．これは，出産年齢に幅があるから[4]で，その結果，例えば図 1 で示したように主個体の生年が同じであるにもかかわらず，ある先祖個体が α にとって 2 世代前だが，β にとって 3 世代前ということもおこる．このように，世代は時間方向の統一した尺度になりにくい．合体過程を時間的な

[3]祖先座標とは可視化のために個体の役割から決めた実数であり，その詳細は以降の議論には本質的には効かない．しかし，気になる人のために以下に定義を書いておこう．ちょうど $G(> 0)$ 世代前の祖先の主個体に対する役割（続柄）は「主個体の」+ $G-1$ 個の「{父もしくは母} の」+「{父もしくは母}」という文字列で表される．この文字列から「の」を省き，主個体を 1，母を 0, 父を 1 で置き換えたものを二進数と見なして得られる自然数を祖先役割番号 A とする．たとえば，A(主個体の母) = 2, A(主個体の父) = 3, A(主個体の母の母) = 4, A(主個体の母の父) = 5 である．また，A(主個体) = 1 と定める．この祖先役割番号 A に対し $x(A) = (2A+1)/2^G - 3$ を祖先役割座標とする．世代 G の祖先役割番号は $2^G \le A \le 2^{G+1} - 1$ を満たすため，祖先役割座標は $-1 + 2^{-G} \le x(A) \le 1 - 2^{-G}$ の範囲をとる．ある祖先の祖先座標は，その祖先が果たしている祖先役割座標の相加平均と定義している．

[4]一生のうちに何度も子供を産むことができることが本質であろう．

側面から特徴付けるため，時間方向の尺度として世代とは別に年代世代という量を定義する．まず，個体 γ の生年を $b(\gamma)$ しよう．すると，生年差 $b(\alpha)-b(\gamma)$ は主個体 α と先祖 γ の時間的な距離を表している．これを平均世代間隔 a で割ったものを，γ の年代世代

$$\tilde{G}_\gamma^\alpha \equiv \left[\frac{b(\alpha)-b(\gamma)}{a}+\frac{1}{2}\right]. \tag{3}$$

とする．ここで，$[x]$ は x を超えない最大の整数であり，平均世代間隔 a は先祖のそれぞれの役割に対して定義される世代間隔 $(b(\alpha)-b(\gamma))/G$ を平均したものである．ここでは，一例として 1996 年に生まれた競走馬のなかからランダムに選ばれた 46 頭を主個体とし，それぞれ 15 世代までさかのぼった木について平均をとった結果 $a \approx 11.47$ となった（図 2 参照）．

この年代世代 $\tilde{G}$ を用いて，二本の木のウェイトベクトル $\{w_\gamma^\alpha\}, \{w_\gamma^\beta\}$ のなす角の余弦 $q^{\alpha,\beta}(\tilde{G})$ の分布を調べたのが図 3 である．年代世代 $\tilde{G}$ に対して前述の 1996 年生まれの 46 個体からランダムに選んだ 100 ペアに対して計算した $q^{\alpha,\beta}(\tilde{G})$ の平均および最大と最小をプロットした．たとえば $q(\tilde{G})$ が 0.5 に達する世代 $\tilde{G}_c$ に着目してみよう．従兄弟のように比較的近いペアが選ばれた場合 $\tilde{G}_c \approx 1$ となるが，血縁的に遠いペアが選ばれると $\tilde{G} \approx 9$ あたりまでかかる．平均的には $\tilde{G} \approx 6$ である．また，12 ～ 13 世代で，二本の木の合体は $\langle q \rangle$ の意味で 99% まで進行していることが分かる．このように，合体初期過程は選択されたペアによる揺らぎが大きいが，後期過程は揺らぎも小さくなっている．

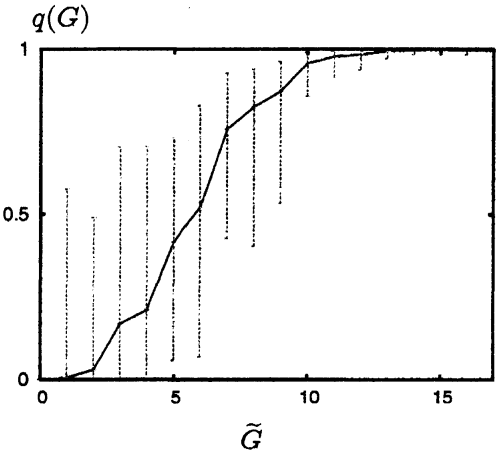

図 3. 二本の木の合体過程．横軸は年代世代．縦軸はウェイトベクトルのなす角の余弦．実線は 1996 年生まれの 46 頭からランダムに選んだ 100 対に関して平均値．破線は，各年代世代における最小から最大を表す．

Derrida らは，1 ペアあたりの平均子数を m とすると，$\langle q(G) \rangle$ が 1% から 99% まで変化するのに要する世代は $\Delta G \approx 4\log_m 10$ であり，$\langle q(G) \rangle = 0.5$ となる世代は $G_c = \log((m-1)N)/\log m - 1$ と見積もっている．競走馬の場合，$m = 2.45$ である [4] ので，$\Delta G \approx 10.3, G_c \approx 11.7$ と予測される．測定によって得られたデータは，前者が $\Delta\tilde{G} \approx 12 \sim 13$ であり，後者が $\tilde{G}_c \approx 6$ ．いずれも，良く一致しているとは言いがたいが，特に後者のずれは大きい．

これらのずれの原因としては，データの有限サイズ効果以外にも様々なものが考えられる．解析に用いられた単純なモデルと実際のデータの間には，既に述べた世代間の重複の有無以外にも，性差の有無や子数の分布などいくつか相違点があることが判明している．これらの相違点が，家系図のネットワーク構造にどう影響を与えるかは今後の課題として残されている．

参考文献

[1] "The Theory of Branching Processes", T. E. Harris, Springer-Verlag, (1963).

[2] S. Ohno, Proc. Nat. Acad. Sci. U.S.A., **93** (1996), 15276–15278.

[3] B. Derrida, S. C. Manrubia, & D. H. Zannette, Phys. Rev. Lett. **82** (1999) 1987–1990; B. Derrida, S. C. Manrubia, & D. H. Zannette, J. theor. Biol. **203** (2000) 303–315.

[4] M. Nishimura and T. Mizuguchi, 数理解析研究所講究録, **1597** (2008) 191–197.

[5] M. Serva, Physica A **332** (2004) 387–393.